Processamento de Linguagem Natural

Atividade de Avaliação Final

Atividade INDIVIDUAL

Entrega: dia da AF

Data da revisão:

Entregáveis: Jupyter notebook contendo o código completo da solução dos exercícios e trechos em markdown explicando o que foi feito.

*Objetivo da atividade: comparar técnicas de classificação diferentes em um problema à sua escolha*

Nesta tarefa, cada aluno deverá escolher uma base de dados com algum tipo de problema de classificação de textos. Se precisar, procure no Kaggle ou no Google por bases de dados para NLP. Nesta tarefa, compararemos diferentes estratégias de classificação para as bases de dados que escolhemos.

O que gostaríamos de descobrir é:

1. Quais são as premissas que fazemos ao tentar classificar esta base?
2. Qual é o desempenho de cada um dos algoritmos de classificação escolhidos na tarefa de classificar essa base?
3. Que tipo de característica (que palavras, estruturas, etc.) cada classificador está mais levando em consideração para fazer a classificação? Isso condiz com o problema em si?
4. Quais elementos

**ENTREGÁVEIS**

Esta atividade deverá ser entregue na forma de um único Jupyter Notebook, através do Blackboard. O notebook deve, obrigatoriamente:

1. Executar sequencialmente (uma vez que o professor vai executar todos os notebooks), sem erros, assumindo as versões e APIs das bibliotecas que usamos nas aulas.
2. Usar o texto em Markdown para explicar claramente todo o raciocínio que foi realizado, acompanhando os códigos em Python. Em especial, devemos ter seções para:
   1. Introdução: com a descrição do problema e a descrição do banco de dados
   2. Desenvolvimento: com uma discussão de todos os classificadores testados, cada um em uma sub-seção.
      1. Cada sub-seção deve ter um título que remeta imediatamente ao classificador que foi testado (por exemplo: “## classificador por votação popular” <- não faça esse teste, porque votação popular não é um classificador prático para NLP).
      2. Todos os classificadores devem vir com uma análise de desempenho e alguma forma de mostrar a relevância de estruturas no processo de classificação
      3. Não se esqueça de evidenciar eventuais processos de limpeza dos dados, incluindo (talvez) a segmentação ou outras etapas de pré-processamento.
   3. Conclusão: retomando os resultados e evidenciando os fatores que você encontrou que são mais relevantes para a classificação, e verificando se algum (ou mais de um) classificador que você fez pode ser usado no problema real relacionado à sua base de dados.

**AVALIAÇÃO**

Os trabalhos serão avaliados de acordo com os seguintes aspectos:

1. (2 pontos) Teste dos classificadores: usamos as metodologias corretas (incluindo separar treino-teste-validação)? As figuras estão claras? Os códigos estão corretos?
2. (2 pontos) Apresentação (texto): os textos estão claros, organizados e concisos? Eles poderiam ser lidos por uma pessoa que não entende de matemática ou de computação (dica: tente escrever o texto para uma pessoa que acabou de ingressar no primeiro ano do seu curso – pense em você no passado)
3. (2 pontos) Apresentação (figuras): os tipos de figuras usados estão adequados à mensagem que vai ser passada? Os eixos têm rótulos claros? O título das figuras guia o leitor em sua interpretação? As cores são consistentes entre as figuras do trabalho (por exemplo: “vermelho” sempre significa “cliente insatisfeito”)? A figura poderia ser destacada do trabalho escrito e, ainda assim, faria sentido?
4. (2 pontos) Conclusões: elas estão baseadas nos dados gerados?
5. (2 pontos) Conclusões: elas ligam os resultados gerados a características da base de dados que você usou (exemplo: “estes dados mostram que músicas sertanejas de raiz têm uma tendência maior a falar de família, fazendas e animais que o hip-hop norte-americano”)

**ANOTAÇÕES IMPORTANTES**

* No dia da revisão para a prova, faremos uma rodada de revisão em pares de textos/figuras em sala. Seu trabalho deve estar completo já nesse dia.
* Este trabalho é INDIVIDUAL, ORIGINAL e AUTORAL. Isso significa que você **não pode** reaproveitar o trabalho de outra disciplina, nem copiar da Internet, nem fazer em grupo.
* Você pode:
  + Encontrar soluções para pequenos problemas de programação usando o Google (por exemplo: “como faço para criar tuplas em Python”) – as documentações existem para serem usadas. Você **não pode** copiar deliberadamente a atividade-fim deste exercício.
  + Discutir com seus colegas, em alto nível, quanto a métodos e resultados que tenha tido. Você **não pode** simplesmente copiar o código ou as frases deles.
  + Procurar o Tiago (pode ser via e-mail – [tiagoft1@insper.edu.br](mailto:tiagoft1@insper.edu.br), nos horários de atendimento, ou passando diretamente na minha sala - TI-25, sétimo andar do prédio 1) para discutir dúvidas metodológicas. Você **não pode** ter uma antecipação da avaliação ou da nota.
* A entrega final é no dia da avaliação final, até o horário em que a aula terminaria. Depois disso, o sistema fechará automaticamente. Não haverá extensão desse prazo uma vez que eu (Tiago) preciso corrigir as avaliações e entregar as notas e o feedback até uma data pré-definida pela Secretaria de Graduação.